Trends in Cognitive Sciences

CellPress OPEN ACCESS

Review

A pull versus push framework for reputation

Jillian J. Jordan[1,*]

Reputation is a powerful driver of human behavior. Reputation systems incentivize 'actors' to take reputation-enhancing actions, and 'evaluators' to reward actors with positive reputations by preferentially cooperating with them. This article proposes a reputation framework that centers the perspective of evaluators by suggesting that reputation systems can create two fundamentally different incentives for evaluators to reward positive reputations. Evaluators may be pulled towards 'good' actors to benefit directly from their reciprocal cooperation, or pushed to cooperate with such actors by normative pressure. I discuss how psychology and behavior might diverge under pull versus push mechanisms, and use this framework to deepen our understanding of the empirical reputation literature and suggest ways that we may better leverage reputation for social good.

Highlights

Reputation has a profound influence on psychology and behavior, and can be leveraged for social good.

In many important contexts, however, we have strong evidence that reputation is functioning, but do not deeply understand how reputation is functioning, from a game-theoretic perspective.

The game-theoretic reputation literature highlights different ways that reputation systems can function. However, these insights have not been well connected to the empirical reputation literature, despite their potential to shed new light on established reputational phenomena, and guide the design of reputation-based interventions.

This article seeks to bridge this gap by proposing a reputation framework that highlights two fundamentally different ways that reputation can function, and by outlining how these mechanisms might give rise to divergent patterns of psychology and behavior.

How reputation shapes human behavior

Humans are deeply concerned with **reputations** (see Glossary) – both their own and those of others. Reputation can motivate prosocial behaviors from charitable giving to environmentalism to diversifying corporate boards [1–10], inspire condemnation and punishment of immoral behavior [11–15], and shape our emotions, cognition, and judgments in ways that facilitate reputation-enhancing actions [12,16,17]. Reputation can also inspire undesired behavior. Reputation can drive people to punish, even when they are unsure that the recipient actually deserves punishment [11,15], and can also fuel discrimination, aggression, cheating, or risky health behaviors, when these actions are socially rewarded [18–23]. Yet while scholars continue to discover new ways that reputation shapes behavior, and draw on reputation-based interventions to encourage desired behavioral changes (e.g., [24–31]), our understanding of reputational phenomena has been limited by a disconnect between the game-theoretic and empirical reputation literatures. Consequently, in many important contexts, we have strong evidence that reputation matters, but do not deeply understand how the underlying **reputation system** is functioning from a game-theoretic perspective.

Here, I suggest that the key to understanding reputation systems is to explain what incentivizes evaluators to reward positive reputations by preferentially **cooperating** with 'good' actors. Drawing on formal models of reputation, I propose a framework that contrasts two fundamentally different incentives: evaluators may be *pulled* towards cooperating with 'good' actors in order to benefit directly from their reciprocal cooperation, or *pushed* to cooperate with such actors by normative pressure. I discuss how psychology and behavior might diverge under the influence of pull versus push mechanisms, and draw on these insights to begin bridging the gap between our theoretical and practical understanding of reputation.

Two mechanisms for reputation

Reputation systems involve 'actors' (who take actions that influence their reputations) and 'evaluators' (who learn about and respond to the behavior of actors). Under stable reputation systems, (i) actors face incentives to take reputation-enhancing actions, and (ii) evaluators face incentives to reward actors with positive reputations by preferentially cooperating with them.

[1] Harvard Business School, Soldiers Field, Boston, MA 02163, USA

*Correspondence: jjordan@hbs.edu (J.J. Jordan).

Game-theoretic models of reputation seek to explain where these incentives come from, and thus why actors and evaluators behave as they do.

If evaluators are more likely to cooperate with actors who have better reputations, actors are straightforwardly incentivized to take reputation-enhancing actions. The key to understanding reputation systems, then, is to explain the incentives of evaluators. What incentivizes evaluators to cooperate with actors who have positive reputations, but not with actors who have negative reputations? Concretely, imagine an interaction between coworkers Andrew (an actor) and Elena (an evaluator). Before their interaction, Andrew can take a reputation-enhancing action (e.g., assisting a colleague). If Andrew knows that having a good reputation will cause Elena to cooperate with him, he will be incentivized to assist his colleague. But what incentivizes Elena to cooperate with Andrew if – and only if – he has a good reputation?

I propose that there are two fundamentally different possible answers to this question, reflecting two fundamentally different mechanisms through which reputation systems can incentivize the rewarding of positive reputations. Under *pull* mechanisms, Elena's incentive to cooperate with a positively regarded Andrew flows from her inference, drawn from his good reputation, that he will likely cooperate with her. This inference can incentivize Elena to cooperate if she is in a **coordination game** context, where she finds it payoff-maximizing to cooperate with Andrew if he cooperates with her. For example, Andrew's good reputation might pull Elena to invest in a collaborative project with him, if investing will pay off for Elena if Andrew works hard and also invests. Conversely, if Andrew has a bad reputation, Elena might infer that investing will not pay off, pulling her away from cooperating.

Under *push* mechanisms, Elena's incentive to cooperate with a positively regarded Andrew instead flows from normative pressure to cooperate with positively regarded community members. When such cooperation is normatively valued (i.e., 'injunctively' normative), if Andrew has a good reputation, cooperating with Andrew can make Elena look good, and declining to cooperate can damage her reputation. For example, Andrew's good reputation might push Elena to help him pack for an overseas move – not because she will directly benefit from helping (packing is effortful, even when the beneficiary is a good person, and Andrew cannot return the favor once he is overseas), but because helping good people reflects positively on Elena. Conversely, if Andrew has a bad reputation, normative pressures will not push Elena to cooperate with him, and could even push her to sanction him.

Importantly, the pull versus push distinction characterizes what drives evaluators such as Elena from an 'ultimate' level of analysis (i.e., what underlying incentives give rise to their behavior). It does not speak to the 'proximate' psychological motives that arise from these incentives. For example, suppose that a push mechanism provides the ultimate incentive for Elena to cooperate with a positively regarded Andrew (i.e., cooperating with Andrew is normatively valued, and will make Elena look good). Although Elena's proximate motive for cooperating could be a conscious desire to appear virtuous, Elena could also be driven by other motives (e.g., the desire to see herself as a good person or to avoid feeling guilty, or her sense that Andrew is a good person who deserves support). Crucially, these proximate psychologies are all compatible with a push mechanism ultimately incentivizing Elena's cooperation.

The theories underlying what I term 'pull' and 'push' mechanisms have long been understood by reputation scholars (e.g., [32–36]), and have sometimes been explicitly contrasted (e.g., [1,37–39]). To date, scholars have used terms such as 'signaling', 'competitive altruism', and 'reputation-based partner choice' to describe what I categorize as pull mechanisms, and

Glossary

Cooperation: acting to confer a benefit on one or more recipients. I conceptualize 'cooperation' broadly to include many ways in which one could confer a benefit. For example, Elena might cooperate with Andrew by helping him to move or by investing in a collaborative project with him. She also might benefit Andrew by declining to impose a cost on him or to enforce a debt or obligation (e.g., as Andrew's manager, Elena might cooperate with Andrew by never asking him to clean the office fridge). Importantly, although cooperation can be costly to the cooperator (e.g., Elena might spend time and effort helping Andrew to move), I also use the term 'cooperation' to encompass acts that benefit the cooperator. For example, Elena might find it payoff-maximizing to invest in a project with Andrew if he also invests; I refer to such situations as 'coordination game contexts'.

Coordination game: an interaction where a focal party finds it payoff-maximizing to cooperate with another party if that party will reciprocally cooperate (and otherwise finds it payoff-maximizing to defect). Thus, the focal party is incentivized to coordinate with (i.e., match the behavior of) the other party.

Defection: declining to cooperate (i.e., not conferring a benefit on one or more recipients), when given the opportunity to do so.

Observable: when a behavior can be witnessed by others, and therefore can influence the reputation of the actor. Importantly, observability is not binary; instead, situations vary with respect to the probability of observation, the number of likely observers, and the importance of these observers to the actor.

Reputation: the set of beliefs or opinions held by others about an actor (e.g., an individual or an organization). Under pull mechanisms, especially important are the beliefs of individual evaluators about the likelihood that an actor will cooperate with them. Under push mechanisms, especially important are public 'reputation states' that reflect a community's consensus about how positively they collectively regard the actor.

Reputation system: a set of interaction structures and behaviors under which actors face stable incentives to engage in behaviors that 'indirect reciprocity' to describe what I categorize as push mechanisms. As outlined below, formal modeling has shown how each type of reputation, through distinct game-theoretic processes, can support stable equilibria in which actors take reputation-enhancing actions and evaluators preferentially cooperate with 'good' actors. In my view, the differences between these two forms of reputation are most clearly understood by contrasting the incentives they create for evaluators to reward positive reputations. Thus, I advance the pull versus push framework to center this distinction and highlight its implications for psychology and behavior. The history of ideas underlying this framework is discussed further in Box 1.

Models of pull mechanisms involve a game with two **stages** (Figure 1A). In stage 1, actors can signal to evaluators that they will cooperate with them in stage 2. In stage 2, evaluators and actors simultaneously decide whether to cooperate with each other. In coordination-game contexts, evaluators can be pulled towards cooperating with actors who have earned 'good' reputations by credibly signaling their cooperativeness in stage 1. Pull mechanisms vary in how actors signal their cooperativeness, and I propose a distinction between signals of 'type' versus 'strategy'. Under type signaling, some 'types' of actors find cooperation payoff-maximizing within stage 2, whereas others find **defection** payoff-maximizing, and actors can signal their type in stage 1. Under strategy signaling, all actors find defection payoff-maximizing within stage 2, but actors can commit to cooperating anyway, and signal their commitment strategy in stage 1. (If all actors find cooperation payoff-maximizing within stage 2, signaling is not needed.) Models of pull mechanisms are discussed further in Box 2.

Models of push mechanisms involve a game that repeats for multiple **rounds** (Figure 1B). In each round, one player (behaving as an evaluator) decides whether to cooperate with another player (behaving as an actor), based on the actor's reputation state (accrued from their behavior in

boost their reputations, and evaluators face stable incentives to reward actors with good reputations by preferentially cooperating with them.

Rounds: iterations, within a game-theoretic model, of a particular type of interaction (i.e., an interaction with a particular structure and set of payoffs for involved parties).

Signals: actions that convey information about some underlying property of an actor. Under pull mechanisms, actors send signals about their likelihood of engaging in cooperative behavior.

Social norms: rules or standards for acceptable behavior, as defined and enforced by a community. Under push mechanisms, social norms codify the reputation consequences of different actions.

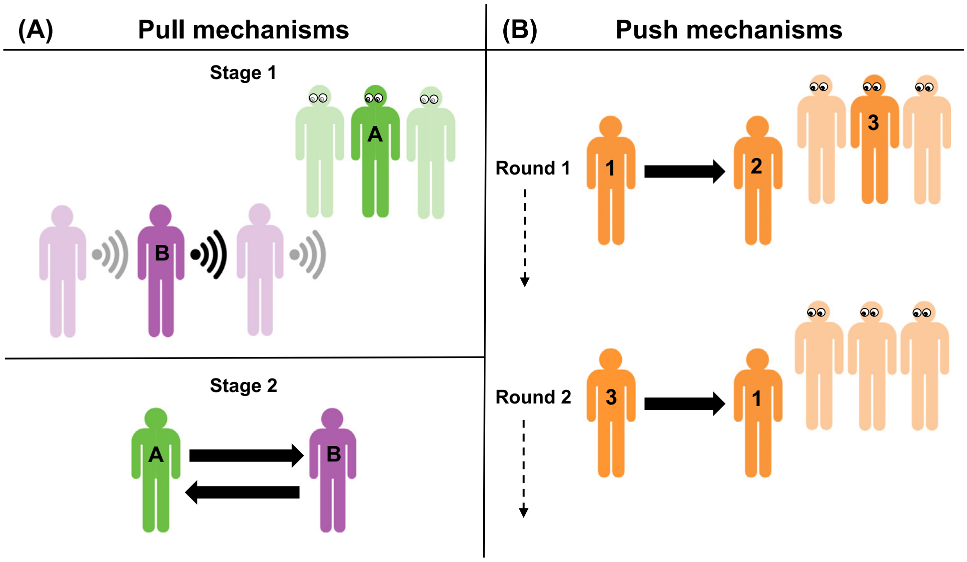
Stages: phases in a game-theoretic model that involve different types of interactions (i.e., interactions with different structures, and payoffs for involved parties).

Box 1. A history of reputation scholarship underlying the pull versus push framework

What I term 'pull' mechanisms have historically been described using terms such as 'signaling', 'competitive altruism', and 'reputation-based partner choice'. Signaling was first formally modeled in 1973 by Spence, who showed that education can serve as an honest indicator of applicant quality in job markets [102]. In 1975, Zahavi tied signaling to the biological sciences [103], prompting scholarship on signaling and cooperation (e.g., [35,36,104]). Much of this work focuses on showing that, when partner choice is possible, individuals who signal cooperativeness can benefit from being selected as partners (over competitors in 'biological markets' [105]) – a process termed 'competitive altruism' [1,2,35] and 'reputation-based partner choice' [37–39]. Broadly, scholarship in this area highlights that reputation-enhancing actions can signal positive information about actors, pulling evaluators toward them.

What I term 'push' mechanisms have been described as 'indirect reciprocity', a term introduced by Alexander in 1987 [33]. The basic idea is that, when an actor cooperates, instead of receiving 'direct' reciprocity from the recipient, he might receive 'indirect' reciprocity from a third-party evaluator. This proposal, first formalized by Boyd and Richardson in 1989 [34], has been modeled in many ways (e.g., [106–108]). These models show that, when an evaluator considers providing indirect reciprocity (i.e., cooperating with an actor who cooperated with somebody else), she must assess not only the prior behavior of the actor (did he cooperate?) but also its normative value (was cooperating appropriate?); the answer bears on her normative obligations towards the actor [107,109]. Broadly, scholarship in this area highlights that reputation-enhancing actions can make actors normatively deserving of support, pushing evaluators to reward them.

Although these reputation mechanisms are each well established, less emphasis has been placed on analyzing their differences [38]. However, scholars have contrasted them (e.g., [1,37–39]), most extensively in a recent article about the theoretical distinctions between reputation-based partner choice and indirect reciprocity [38] (a comparison that maps closely onto the pull versus push distinction, although I argue that both pull and push mechanisms can operate with and without partner choice). Still, our theoretical understanding of the different ways that reputation can function has not been well connected to empirical reputation scholarship in psychology, cognitive science, and behavioral science. The pull versus push framework aims to bridge this gap by spotlighting a fundamental difference between the above-described categories of reputation – the incentives they create for evaluators to reward positive reputations – and highlighting the relevance of this distinction for psychology and behavior.



Trends in Cognitive Sciences

Figure 1. Two mechanisms for reputation. (A) Pull mechanisms involve a two-stage game with two roles: player A is the evaluator, and player B is the actor. In stage 1, player A observes while player B sends signals about how he will behave in stage 2. In stage 2, players A and B decide whether to cooperate with each other. Player A finds cooperation payoff-maximizing if, and only if, player B cooperates. This 'coordination game' structure can pull player A to cooperate if player B has credibly signaled in stage 1 that he will cooperate in stage 2. When partner choice is possible, player A can choose which player B to interact with in stage 2 – and can base this decision, in addition to her subsequent cooperation decision, on the signals that player Bs send in stage 1. (B) Push mechanisms involve a game that repeats for multiple rounds. Players have public reputation states (e.g., good or bad), and social norms describe how reputations are determined (e.g., to remain good, cooperate with good co-players but not with bad co-players). The game has only one role, so in a given round players may behave as an evaluator or an actor. In each round, a player behaving as an evaluator decides whether to cooperate with a player behaving as an actor, based on the actor's reputation. The evaluator's reputation then updates according to the norm. Even if players never find cooperating payoff-maximizing within a round, norms can push evaluators to cooperate with 'good' actors in order to receive cooperation from others evaluating them in future rounds. Thus, if player 2 has a good reputation, player 1 might cooperate with him in round 1 so that in round 2, she will have a good reputation and player 3 will cooperate with her. When partner choice is possible, evaluators can choose which actors to interact with – and can base such decisions, in addition to their cooperation decisions, on the reputations of actors.

previous rounds). Evaluators can be pushed to cooperate with actors who have positive reputations when such cooperation is normatively valued (and will thus benefit their own reputations, making them more likely to receive cooperation from others evaluating them in future rounds). In particular, **social norms** must stipulate that (i) there are reputation benefits for cooperating with co-players with positive reputations, and (ii) the reputation consequences of cooperating are less positive when the reputation of one's co-player is less positive. If reputation states are binary (good or bad), cooperating with 'good' co-players must be required for a good reputation, whereas cooperating with 'bad' co-players must not be required (and may even induce a bad reputation). Norms can also place value on other actions beyond cooperation with 'good' co-players (e.g., public goods contributions). Models of push mechanisms are discussed further in Box 3.

Contrasting the game-theoretic processes underlying pull versus push mechanisms highlights that the mechanisms rely on distinct incentive and interaction structures, with implications for the contexts in which we should expect to observe each form of reputation. For example, we should anticipate pull mechanisms in situations with coordination-game structures, and push mechanisms in situations with well-defined social norms.

Box 2. Models of pull mechanisms

Some models of pull mechanisms (e.g., [7,32,49,101,110]) involve type signaling. In these models, actors (i.e., player Bs in Figure 1) vary in their incentives: some 'types' find cooperation payoff-maximizing within stage 2, whereas others find defection payoff-maximizing. In stage 2, actors act to maximize their payoffs; thus, an individual's type determines whether he cooperates. And in stage 1, actors signal their type to evaluators (i.e., player As).

How do differences in type arise? Type-signaling models assume that processes outside pull-based reputation (e.g., direct reciprocity, institutional reward or punishment, or push-based reputation) incentivize some actors to cooperate. Within this framework, actors might vary in their skills or resources (and thus their costs of cooperating) and/or exposure to these incentivizing processes (and thus the benefits they receive from cooperating) such that some (but not all) find cooperation net-positive and are cooperative types.

To signal their type, cooperative actors can send 'costly signals': actions that are less costly (and/or more beneficial) for cooperative types [32,36,103]. For signals to stay honest, uncooperative types must find signaling too net-costly to be worthwhile – even when considering the benefit of (falsely) appearing cooperative to evaluators.

Most straightforwardly, cooperating today can signal future cooperation [111]. If Andrew assists his colleague, Elena may infer that Andrew finds producing good work easy (e.g., because he is skilled) and/or beneficial (e.g., because he anticipates staying at their workplace long enough to reap the rewards of building relationships), and therefore will cooperate (e.g., by investigating in a collaboration) with her.

The literature also highlights specific actions that may serve as costly signals [7,49,50,101,105,110,112–119] – either of lower cooperation costs (e.g., hunters may signal skills by targeting high-risk high-reward prey [112]; wealthy people may signal resources via conspicuous philanthropy [113]) or greater benefits from cooperating (e.g., helping a close relationship partner may signal that you benefit from their success [110]). Signals may also convey an intention to cooperate in the future [110,120], reflecting the perception that cooperation is net-beneficial (e.g., costly courtship rituals, such as engagement rings, are not worthwhile for suitors who will soon end a relationship, and costly apologies are not worthwhile for transgressors who will soon transgress again; such actions thus signal cooperative intent [116–118]).

Other models of pull mechanisms (e.g., [87,121,122]) involve strategy signaling. Here, all actors face the same incentives, and find defection payoff-maximizing within stage 2. However, in stage 1, actors can adopt a strategy that pre-commits them to cooperating anyway. In stage 2, actors follow their predetermined strategy; thus, the strategy adopted by an individual determines whether he cooperates. And in stage 1, actors signal their strategy to evaluators. For strategies to be credible signals, they must be difficult to fake and reliably cause cooperation.

One way that actors may signal a commitment to cooperating is through cooperation-inducing emotions, values, or habits. In his theory of 'strategic emotions', Frank proposes that by signaling emotions (e.g., love) and values (e.g., integrity) that compel cooperation, even when defecting would be payoff-maximizing, people can elicit trust from others [16,121]. For example, if Andrew is passionate about work, guilt-prone [123], or habitually works long hours, Elena might trust him to invest in a collaboration with her – even if he would, materially, be better off slacking.

Another strategy signal may be to decline to consider the costs of cooperation. Imagine that you typically find cooperation payoff-maximizing, but cooperating is sometimes sufficiently costly that it is not net-beneficial. A model by Hoffman et al. [87] reveals that, in this scenario, declining to evaluate the cooperation cost in each instance can signal that you will cooperate reliably, because you will not know when defecting would serve you better. Consequently, cooperating 'without looking' can signal trustworthiness [87,124].

Importantly, pull and push mechanisms can both operate in contexts where 'partner choice' is, or is not, possible (i.e., where evaluators can, or cannot, choose which actors to interact with). When partner choice is possible, evaluators can use the reputations of actors to decide not merely whether to cooperate with them in a given interaction, but also whether to interact with them in the first place. Thus, under pull mechanisms, if an actor signals his cooperativeness in stage 1, an evaluator might choose to interact (and then cooperate) with him in stage 2. This evaluator would be pulled towards a 'good' interaction partner to benefit from his cooperation. Under push mechanisms, an evaluator might interact (and cooperate) with a 'good' actor in round 1, to be interacted (and cooperated) with by a different evaluator in round 2. Here, evaluators are pushed towards 'good' partners by norms governing whom one should partner with (e.g., norms to interact and cooperate with 'good' individuals, but ostracize 'bad' individuals).

Box 3. Models of push mechanisms

Models of push mechanisms, historically termed 'indirect reciprocity', take various forms (e.g., [34,106–108,125]). The simplest model is image scoring, where the reputation of a player reflects the number of times he has cooperated [106]. Players can condition cooperation on the image scores of co-players, allowing 'prosocial discriminators' who cooperate only with co-players with high scores. If prosocial discriminators are prevalent, players face incentives to cooperate in order to earn high scores and receive cooperation.

Yet this model cannot explain why prosocial discriminators should discriminate. Because the image score of a player tracks their cooperation with any co-player, players who indiscriminately cooperate will have better reputations than prosocial discriminators. And without prosocial discriminators, there is no incentive to have a high image score. Thus, image scoring cannot support stable cooperation [107,125,126] because it cannot answer the key question this article centers: what incentivizes evaluators to cooperate preferentially with actors who have positive reputations?

A satisfying answer comes from 'standing' models [34,107–109]. In these models, reputation states are binary, and to maintain 'good standing' (i.e., a good reputation), players are only required to cooperate with others in good standing. Thus, the social norm requires cooperating with 'good' co-players, but allows 'justified defection' against 'bad' co-players. (Such defection may even be required, such that players who cooperate with bad co-players become bad themselves [108].) Under standing, prosocial discriminators outperform non-discriminators, who incur the cost of cooperating with bad co-players at no reputation benefit (or a reputation cost). Thus, standing creates stable incentives for evaluators to reward positive reputations, and supports stable cooperation [107,108,126–128]. Recently, scholars have begun investigating how the logic of standing may extend when reputation states are not binary (i.e., there are more than two levels of reputation standing) [129,130].

Importantly, norms can also prescribe actions beyond cooperation with 'good' co-players (e.g., public goods contributions, participation in rituals, derogation of out-group members) [131]. Push mechanisms can thus incentivize any action that is normatively valued – including actions that are not socially beneficial – provided that the costs of adhering to the norm are outweighed by the reputational benefits [73]. Norms can also vary in the consequences they create for good and bad reputations. For example, in addition to requiring cooperation with 'good' co-players, norms may require punishment of 'bad' co-players [73,132,133]. There are thus infinite potentially stable norms. Norms that benefit groups, or that benefit individuals with coordinating power over groups, may be favored by equilibrium-selection processes [134,135].

Bridging the gap from theory to practice

The empirical reputation literature highlights a growing list of actions that are reputationally rewarded by evaluators, including behaving prosocially [1,7,39–41], punishing wrongdoers [14,15,42–49], making emotional or uncalculating cooperative decisions [50–53], making deontological (vs. utilitarian) moral decisions [54], taking stances on controversial issues [55,56], and displaying decision-making 'biases' such as escalating commitment [57,58]. Moreover, we know that actors preferentially enact many of these reputationally beneficial actions when others are watching [3,5,7,8,11,13,15,24,50] and can reward them [1,2,5,7].

When considering this work through the pull versus push framework, it becomes clear that the interpretation of these findings may depend meaningfully on the underlying reputation mechanism(s). For example, companies frequently broadcast prosocially minded initiatives (e.g., charitable donations, corporate social responsibility), and consumers and workers reward such initiatives by preferentially patronizing, and choosing to work for, prosocial companies [59–64]. In any particular instance, these phenomena could reflect that consumers and workers are *pulled* towards prosocial companies because they anticipate better treatment from them (e.g., more reliable products, customer service, or warranties as consumers, or equitable and generous working conditions as employees) [65,66]. Alternatively, consumers and workers might be *pushed* towards prosocial companies by normative pressure (reflecting that patronizing and working for such companies looks good in the eyes of others) [64]. Or, both mechanisms could be at play. Of course, the pull versus push distinction also shapes the interpretation of actors' behavior – when companies advertise prosocial initiatives, are they signaling cooperativeness or adhering to norms?

In many reputational contexts, however, we know little about how reputation operates at an ultimate level, reflecting that our game-theoretic understanding of how reputation can function has not been well connected to the empirical reputation literature. In the sections that follow, I draw on the pull versus push framework to begin to bridge this gap. I outline ways that psychology and behavior might diverge under pull versus push mechanisms, and discuss how these predictions may illuminate how reputation is functioning in contexts of interest, highlight directions for future inquiry, and deepen our understanding of established empirical findings. I then leverage these insights to offer suggestions for the design of reputation-based interventions.

Divergent predictions of pull versus push mechanisms

Are evaluators concerned with their own reputations?

When evaluators assess and reward the reputations of others, are they driven – consciously or unconsciously – by concerns with their own reputations? Under pull mechanisms, evaluators directly benefit from cooperating with 'good' actors, so we might expect the answer to be no. Under push mechanisms, however, evaluators benefit from rewarding 'good' actors by themselves looking good, so reputation concerns might mediate this rewarding.

As such, in contexts of interest, the efforts of evaluators to manage their own reputations might speak to the influence of push mechanisms. For example, if consumers become more sensitive to the charitable records of companies when making purchases that are public, we might conclude that normative pressures push consumers to support charitable companies. (If consumers are similarly sensitive to charitable records in private versus public, however, the inference might be more ambiguous. Consumers could primarily be pulled towards charitable companies, or they could be pushed by norms but nonetheless remain strongly inclined to reward charitable companies even in private – for example, because they have internalized the norm for such rewarding and/or heuristically assume that reputation is typically at stake [12].)

Thus, the pull versus push framework highlights an interesting future direction: investigating when and how concerns with our own reputations shape our responsivity to the reputations of others.

Do evaluators prioritize first-order beliefs or higher-order beliefs about reputation?

Are evaluators more concerned with the reputations of actors in their own eyes, or in the eyes of others? For example, imagine that a company is publicly regarded as being very charitable, but a consumer has private information that its charitable efforts are greatly exaggerated. When deciding whether to patronize the company, will she prioritize her first-order beliefs about the company or her higher-order beliefs about how others view the company?

Under pull mechanisms (which incentivize cooperation with actors that one personally trusts), we might expect evaluators to prioritize first-order beliefs. By contrast, under push mechanisms (which incentivize cooperation with actors that one's community approves of [67]), evaluators might be concerned with higher-order beliefs and whether reputation information is 'common knowledge' (i.e., is publicly in the open). Thus, in contexts of interest, the sensitivity of evaluators to their private impressions versus the public reputations of actors might illuminate the role of pull versus push mechanisms.

Through this insight, the pull versus push framework may help to explain several interesting reputational phenomena involving sensitivity to higher-order beliefs and common knowledge [68]. For example, actors anticipate feeling less embarrassment and shame when they imagine committing transgressions that are witnessed by others, but do not become common knowledge [17]. Moreover, actors sometimes use indirect speech when initiating transgressions (e.g., 'Officer, is there some way we can take care of the ticket?' to propose a bribe [69]) to create 'plausible deniability' and avoid making their violation common knowledge [70]. Push mechanisms may help to explain these phenomenona: even if an evaluator knows that an actor has transgressed, if the transgression is not common knowledge (and thus does not damage the public reputation of the actor), normative pressures may push the evaluator to continue behaving as if the actor is 'good'.

By a similar logic, push mechanisms may help to explain why unpopular norms can persist. Evidence about 'pluralistic ignorance' highlights norms that people privately dislike, but mistakenly believe that others approve of (e.g., norms for heavy drinking on American college campuses [23] or against women working in Saudi Arabia [71]). Push mechanisms may contribute to the enforcement [72], and thus the stability, of such norms: even if an evaluator personally dislikes a norm, normative pressures may push her to reward norm-following actors, given their positive public reputations.

What does it mean to be 'good' or 'moral'?

How should evaluators assess whether an actor is 'good' or 'moral'? Under pull mechanisms (under which evaluators cooperate with 'moral' actors to directly benefit from their good treatment), we might expect evaluators to attend specifically to signals of an actor's proclivity to cooperate with them. By contrast, under push mechanisms (under which evaluators cooperate with 'moral' actors to meet normative obligations), evaluators might take a broader viewpoint, seeing actors as 'moral' whenever they are viewed, by society, as deserving positive treatment. Thus, in contexts of interest, the breadth of evaluators' moral conceptions – and actors' efforts to build their moral reputations – may speak to the influence of pull and push mechanisms.

Through this insight, the pull versus push framework may help to illuminate many interesting ways that our morality can extend broadly. For example, some consumers avoid patronizing Amazon because they see the company as immoral on account of its treatment of *workers* – even while Amazon arguably treats its *customers* uniquely well (e.g., by providing low prices, fast shipping, and flexible returns). Push mechanisms provide an ultimate explanation for such phenomena, insofar as boycotting Amazon is normatively rewarded in some communities.

Moreover, push mechanisms may help to explain even broader moral conceptions. Although our normative obligations to an actor can depend on his moral conduct (towards us or others), they can also depend on other factors [73] (e.g., does he perform cultural rituals? Is he a victim of wrongdoing?). As such, push mechanisms can help to explain why our conceptions of 'morality' are sometimes sensitive to factors that shape our normative obligations towards others, but do not correlate with their moral conduct.

For example, people sometimes see victims of wrongdoing as more morally virtuous than non-victims – despite not actually believing that victims behave especially morally [74]. This 'virtuous victim effect' raises a puzzle: why should somebody seem more moral merely because they were mistreated by somebody else? Evidence supports the explanation that normative pressures push us to support victims, making it beneficial to see them as moral (and therefore deserving of support), even if they are not particularly inclined towards moral behavior [74].

What factors can discredit the reputation value of an action?

When evaluators assess the behavior of an actor, what factors can strip a potentially reputation-enhancing act of its reputation value? Under pull mechanisms (which incentivize evaluators to assess whether actors will actually treat them well), evaluators may be concerned with 'false' signaling (i.e., the possibility that an act inaccurately implies that the actor is cooperative). By contrast, under push mechanisms (which incentivize evaluators to assess the normative value of an actor's behavior), evaluators may be concerned with whether an act is directed at an inappropriate recipient – given that that the same prosocial act might be normatively valued when directed at a cooperator, in-group member, or high-status individual, but frowned upon when directed at a transgressor, out-group member, or low-status individual [75,76]. Thus, in contexts of interest, the ways in which evaluators discount the seemingly prosocial behavior of actors, and actors protect against such discounting, may illuminate the role of pull and push mechanisms.

In this way, the pull versus push framework can help us to understand several interesting dynamics surrounding this discounting. For example, evaluators sometimes discount prosocial acts that were reputationally motivated [2,77,78, 79–82] and may therefore seem 'tainted' [83]. In one study, subjects were less likely to trust someone who behaved generously in a 'dictator game' in a subsequent 'trust game' if the dictator knew that the trust game was coming [79]. Moreover, terms such as 'virtue signaling' [84], 'moral grandstanding' [85], and 'slacktivism' are increasingly used to accuse do-gooders of being merely reputationally motivated [86]. Pull mechanisms can help to explain these observations, insofar as reputationally motivated prosociality provides a weaker signal of cooperativeness.

This insight also highlights one reason that actors may make cooperative decisions that are quick and uncalculating [50,87], and thus not strongly sensitive to reputational considerations [12]. Uncalculating decision-makers risk cooperating when nobody is watching, and thus incurring the costs of cooperation with no reputational upside. However, because uncalculating cooperation appears to be less conditional on expected reputational gains, it is an especially strong signal – so uncalculating actors can earn larger reputational benefits for cooperating if they are observed [50–53].

Moreover, pull mechanisms, and the concerns with false signaling they engender, can help to explain why we dislike hypocrites. Why can moral acts that normally enhance the reputations of actors (e.g., condemning wrongdoers, encouraging prosocial behavior) seem hypocritical, and actually harm an actor's reputation, when paired with an inconsistent moral track record [48,88–95]? A key reason may be that, when paired with transgressive behavior, moral engagement can constitute false signaling (by implying that the actor does not, in fact, transgress) [48,88,96]. Supporting this hypothesis, evaluators forgive hypocrites who avoid false signaling by openly confessing their transgressions (e.g., 'I think it is wrong to download music illegally, but I sometimes do it anyway') [48].

Push mechanisms, by contrast, may help to explain why the reputation value of prosocial behavior can diminish or reverse when directed at 'bad' recipients. For example, although corporate donations are normally seen positively, some liberal consumers boycotted Chik-fil-A following their donations to Christian charities that opposed same-sex marriage. Similarly, experimental subjects are less likely to cooperate with individuals who have cooperated with defectors than with cooperators [97]. And research on 'second-order punishment' shows that cooperating with transgressors can harm one's reputation: people who cooperate with wrongdoers – or simply fail to punish them – may themselves be punished [98,99]. Pull mechanisms can help to explain these phenomena, insofar as the normative value of prosocial acts depends on their recipients.

Implications for reputation-based interventions

Mounting evidence highlights the potential for interventions to leverage reputation to encourage desired behaviors (e.g., [24–31]). How might the pull versus push framework inform such efforts?

Table 1. Key table. Distinctions between pull and push mechanisms[a]

		Pull mechanisms	Push mechanisms
Game-theoretic processes	What incentivizes evaluators to reward 'good' actors?	Evaluators cooperate with 'good' actors (i.e., actors with positive reputations) to directly benefit from their reciprocal cooperation.	Evaluators cooperate with 'good' actors because such cooperation is normatively valued. When an evaluator cooperates with a 'good' actor, it reflects positively on the reputation of the evaluator, making it more likely that she will receive cooperation from others evaluating her in the future.
	What defines a 'good' reputation?	To earn a 'good' reputation, an actor must credibly signal to an evaluator that he will cooperate with her in the future. Thus, a good reputation requires (only) actions that convey positive information about the proclivity of an actor to cooperate with a relevant evaluator.	To earn a 'good' reputation, individuals (who can behave, in a given interaction, as an actor or evaluator) must take actions that are normatively valued by their community. For push mechanisms to stabilize cooperation, cooperating with 'good' actors must be normatively valued – and valued more than cooperating with 'bad' actors. Norms can also place value on other actions, which may be prosocial (e.g., public goods contributions), neutral (e.g., participation in arbitrary rituals), or even value-destructive (e.g., aggression or discrimination).
	How must interactions be structured for the reputation system to function?	A 'coordination game' structure is needed. Evaluators must find it payoff-maximizing to cooperate with an actor, if (and only if) that actor will also cooperate. This reflects that evaluators benefit directly from cooperating with 'good' actors (who they expect to cooperate with them), within their dyadic interactions. Thus, for evaluators to face incentives to cooperate with 'good' actors, it is not necessary that (i) they are currently being observed, or (ii) others will evaluate them in the future.	No 'coordination game' structure is needed. It is not necessary that evaluators ever find it payoff-maximizing to cooperate with actors, within their dyadic interactions. This reflects that evaluators benefit from cooperating with 'good' actors in subsequent interactions (in which others are evaluating them). Thus, for evaluators to face incentives to cooperate with 'good' actors, it must be sufficiently likely that (i) they are currently being observed, and (ii) others will evaluate them in the future.
Predictions for psychology and behavior	Are evaluators concerned with their own reputations?	When evaluating and responding to the reputations of others (and deciding whether to cooperate with 'good' actors), evaluators do not need to be concerned with their own reputations.	When evaluating and responding to the reputations of others (and deciding whether to cooperate with 'good' actors), evaluators may be concerned – consciously or unconsciously – with maintaining their own reputations.
	Do evaluators prioritize first-order beliefs or higher-order beliefs?	Evaluators may be primarily concerned with their own personal perceptions of the reputations of actors (i.e., first-order beliefs about reputation information).	Evaluators may be concerned with others' perceptions of the reputations of actors (i.e., higher-order beliefs about reputation information), and whether reputation information is common knowledge.
	How do evaluators conceptualize 'good' or 'moral'?	Evaluators may attend specifically to signals of whether an actor will behave cooperatively towards them.	Evaluators may take a broader viewpoint, and see actors as 'good' or 'moral' whenever they are viewed, by society, as deserving positive treatment. Evaluators' moral conceptions may therefore depend on a wide range of factors (e.g., whether an actor keeps kosher, or is a victim).
	What discredits the reputation value of an action?	Evaluators may discount seemingly prosocial acts if they perceive evidence of false signaling, which might give rise to (i) discounting of reputationally motivated prosociality, (ii) accusations of 'virtue signaling', 'moral grandstanding', or 'slacktivism', and (iii) negative evaluations or punishment of perceived hypocrisy.	Evaluators may discount seemingly prosocial acts if they are directed at the 'wrong' (i.e., normatively inappropriate) recipients, which might give rise to negative evaluations or second-order punishment of prosocial behavior towards transgressors, out--group members, or low-status individuals.

Table 1. (continued)

		Pull mechanisms	Push mechanisms
Considerations for reputation-based interventions	When is reputation relevant?	To assess whether a reputation-based intervention could promote a desired behavior, we should ask: could the desired behavior be a credible signal of cooperativeness?	To assess whether a reputation-based intervention could promote a desired behavior, we should ask: could the desired behavior be normatively valued?
	Which behaviors should be made more observable?	It might be sufficient to make the desired actor behavior more observable. It might not be a priority to make the behavior of evaluators more observable, given that evaluators benefit directly from cooperating with actors who engage in the desired behavior.	It might not be sufficient to make the desired actor behavior more observable. We might also wish to make the behavior of evaluators more observable, given that evaluators benefit reputationally from cooperating with actors who engage in the desired behavior.
	Other considerations?	When encouraging a desired actor behavior, we might try to help actors to avoid being seen as false signalers.	To encourage evaluators to reward 'good' actors, we might try to help the community reach consensus (i.e., common knowledge) about which actors are 'good'.

[a] This table (i) contrasts the game-theoretic processes underlying pull versus push mechanisms, and conditions under which each mechanism can operate; (ii) highlights the potentially divergent psychological and behavioral predictions of pull versus push mechanisms; and (iii) suggests considerations for reputation-based interventions in contexts where pull versus push mechanisms are operative.

To begin, the framework suggests two questions to ask when assessing the potential for reputation to promote a desired behavior, and thus the relevance of reputation as an intervention strategy. First, regarding pull mechanisms: does the desired behavior signal cooperativeness to a relevant evaluator audience? For example, if we wish to encourage companies to make charitable donations, do consumers or workers anticipate better treatment from charitable companies? Second, regarding push mechanisms: is the desired behavior normatively valued? For example, do people see charitable companies as being normatively deserving of support? If at least one answer is 'yes', a reputation-based intervention might be effective.

Furthermore, the pull versus push framework suggests ways we might tailor reputation-based interventions to the operative reputation mechanism(s) in our context. For example, reputation-based interventions often make desired behaviors more **observable** to increase their reputational benefits [24,27–31]. But which actions should we shine the spotlight on? Under both pull and push mechanisms, it may prove fruitful to make the desired actor behavior (e.g., corporate donations) more observable. Might it also prove fruitful, however, to make the behavior of evaluators rewarding that behavior (e.g., consumers or workers) more observable? Amplifying the observability of evaluators' behavior may have limited value if pull mechanisms are predominant (and consumers and workers are primarily driven towards charitable companies because they anticipate positive treatment from them). If push mechanisms are operative, however, it might be valuable to help consumers and workers to broadcast the charitable companies they patronize and work for so that they can benefit reputationally from these affiliations.

Moreover, when push mechanisms are operative, it may be helpful to facilitate public consensus (i.e., common knowledge) about who has a good reputation, to give people confidence that others will reward their cooperation with 'good' actors. For example, to encourage consumers to patronize charitable companies, we might help to align views about who is 'charitable' (e.g., by publicly ranking the charitable records of companies or by encouraging public discussion of such records on social media). When pull mechanisms are operative, it may be fruitful to encourage versions of desired behaviors that will seem to be credible indicators of virtue, even when paired with an imperfect moral track record (to avoid accusations of false signaling or hypocrisy). To this end, we might nudge companies to make charitable donations that are costly, efficacious, and paired with acknowledgments of any relevant moral shortcomings in the company's history [88].

Importantly, these insights also apply to reputation-based interventions aimed at discouraging undesired behavior. For example, imagine an intervention that makes corporate pollution more observable so as to heighten its reputational costs. If normative pressures push consumers and workers to avoid high-pollution companies, we might also make the decisions of these evaluators more observable (to heighten the reputation costs of affiliating with high-pollution companies). Alternatively, imagine that we wish to disrupt an undesired norm that is actually unpopular (e.g., heavy college drinking). If norms push students to enforce the norm by rewarding heavy drinkers, we might aim to facilitate public consensus that the norm is disliked (so that heavy drinkers will stop having positive public reputations) – for example, by broadcasting statistics about attitudes towards drinking, and encouraging public discussion of these statistics [100].

Concluding remarks

Humans care deeply about reputations, as illustrated by two stylized facts. First, actors engage in reputation-enhancing actions so as to be seen positively by evaluators. Second, evaluators reinforce positive reputations by behaving more cooperatively towards 'good' actors than 'bad' actors. The first fact is readily explained by the second, so the key to understanding reputation systems is to uncover the incentives of evaluators.

This article has contrasted two fundamentally different incentives that evaluators may face. First, evaluators may be *pulled* towards good actors (and away from bad actors) so as to benefit directly from the cooperative conduct of good actors. Second, evaluators may be *pushed* to cooperate with good actors (but not bad actors) by normative pressure. Pull versus push mechanisms have distinct game-theoretic underpinnings (i.e., they rely on different incentive and interaction structures), and therefore may operate in different contexts and give rise to divergent patterns of psychology and behavior, with implications for reputation-based interventions. By highlighting these differences, summarized in Table 1, the pull versus push framework aims to help bridge the gap between our understanding of reputation in theory and practice.

The framework also suggests several directions for future inquiry (see Outstanding questions) to further this agenda. In particular, scholars might investigate how pull and push mechanisms work together to shape psychology and behavior. Although reputation theorists have historically modeled these mechanisms independently, and this article has contrasted them, they are not mutually exclusive – raising the question of how pull and push mechanisms influence and reinforce each other. For instance, might actors advertise their cooperativeness (under pull-based reputation) by broadcasting their push-based incentives to adhere to cooperative norms [101] (e.g., by signaling that they have internalized such norms)? Furthermore, although this article has focused on why evaluators cooperate with good actors, evaluators also reinforce positive reputations by punishing bad actors, and draw a rich array of inferences about actors that guide their treatment of them. Future work might investigate these mediating inferences (and how they differ across mechanisms) and explore the utility of contrasting pull versus push incentives for punishment.

Acknowledgments

The author gratefully acknowledges very helpful feedback and discussion from Nathan Barker, Adam Bear, Robert Boyd, Diane Brinkley, Charles Dorison, Amit Goldenberg, Nicholas Jordan, Nour Kteily, Martin Nowak, David Rand, Michael Stagnaro, and Emmet Sullivan, as well as the Editor, Lindsey Drayton, and three anonymous reviewers.

Declaration of interests

The author declares no conflicts of interest.

Outstanding questions

When evaluators are pulled versus pushed to cooperate with 'good' actors, what mediating inferences (e.g., attributions of traits such as authenticity or loyalty) guide their decisions? A complete investigation might look beyond inferences directly related to an actor's cooperativeness. For example, if an actor seems to be competent, do evaluators infer that he can easily provide benefits, pulling them to cooperate? Alternatively, does high status or charisma imply good standing, pushing evaluators to cooperate?

Evaluators reinforce positive reputations not merely by cooperating with 'good' actors but also by punishing 'bad' actors. Might the pull versus push framework shed light on punishment? Indeed, norms for punishment can *push* evaluators to punish, and evaluators might be *pulled* towards punishing when they can benefit directly from deterring future transgressions. How might these distinct punishment incentives shape psychology and behavior?

Both pull and push mechanisms are clearly important – but we know little about their relative prevalence and power, or about how often they operate independently versus co-occur. How frequently are reputational phenomena driven by pull mechanisms, push mechanisms, or both mechanisms? And do the mechanisms often 'spill over' to shape behavior, even when the underlying incentives are absent?

When push and pull mechanisms co-occur, do they interact and reinforce each other? For example, pull mechanisms incentivize actors to signal cooperativeness; could this be achieved by signaling a proclivity, supported by push mechanisms, to follow cooperative norms (e.g., by signaling that one has internalized such norms)? Alternatively, might norms against deception push people to punish 'false signalers' – bolstering the stability of pull mechanisms by helping to keep signals honest?

How might the expressions or importance of pull and push mechanisms vary across individuals (e.g., as a function of personality), cultures (e.g., tight vs loose, individualist vs collectivist), or development (e.g., do children become sensitive to one mechanism before the other)?

References

1. Barclay, P. (2004) Trustworthiness and competitive altruism can also solve the 'tragedy of the commons'. *Evol. Hum. Behav.* 25, 209–220
2. Barclay, P. and Willer, R. (2007) Partner choice creates competitive altruism in humans. *Proc. Biol. Sci.* 274, 749–753
3. Engelmann, D. and Fischbacher, U. (2009) Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav.* 67, 399–407
4. Feinberg, M. *et al.* (2014) Gossip and ostracism promote cooperation in groups. *Psychol. Sci.* 25, 656–664
5. Herrmann, E. *et al.* (2019) Children engage in competitive altruism. *J. Exp. Child Psychol.* 179, 176–189
6. Wu, J. *et al.* (2016) Gossip versus punishment: the efficiency of reputation to promote and maintain cooperation. *Sci. Rep.* 6, 23919
7. Barclay, P. and Barker, J.L. (2020) Greener than thou: people who protect the environment are more cooperative, compete to be environmental, and benefit from reputation. *J. Environ. Psychol.* 72, 101441
8. Engelmann, J.M. *et al.* (2018) Concern for group reputation increases prosociality in young children. *Psychol. Sci.* 29, 181–190
9. Chang, E.H. *et al.* (2019) Diversity thresholds: how social norms, visibility, and scrutiny relate to group composition. *Acad. Manag. J.* 62, 144–171
10. Harbaugh, W.T. (1998) The prestige motive for making charitable transfers. *Am. Econ. Rev.* 88, 277–282
11. Jordan, J. and Kteily, N. (2020) Reputation fuels moralistic punishment that people judge to be questionably merited. *PsyArXiv* Published online March 21, 2020. https://psyarxiv.com/97nhj
12. Jordan, J.J. and Rand, D.G. (2020) Signaling when nobody is watching: a reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *J. Pers. Soc. Psychol.* 118, 57–88
13. Kurzban, R. *et al.* (2007) Audience effects on moralistic punishment. *Evol. Hum. Behav.* 28, 75–84
14. Batistoni, T. *et al.* (2022) Third-party punishers do not compete to be chosen as partners in an experimental game. *Proc. R. Soc. B* 289, 20211773
15. Jordan, J. and Kteily, N. (2022) How reputation does (and does not) drive people to punish without looking. *Proc. Natl. Acad. Sci. U. S. A.* (in press)
16. Frank, R.H. (2011) The strategic role of the emotions. *Emot. Rev.* 3, 252–254
17. Thomas, K.A. *et al.* (2018) Common knowledge, coordination, and the logic of self-conscious emotions. *Evol. Hum. Behav.* 39, 179–190
18. Nawata, K. (2020) A glorious warrior in war: cross-cultural evidence of honor culture, social rewards for warriors, and intergroup conflict. *Group Process. Intergr. Relat.* 23, 598–611
19. Nisbett, R.E. (1993) Violence and US regional culture. *Am. Psychol.* 48, 441
20. Whitaker, R.M. *et al.* (2018) Indirect reciprocity and the evolution of prejudicial groups. *Sci. Rep.* 8, 13247
21. Zhao, L. *et al.* (2018) Telling young children they have a reputation for being smart promotes cheating. *Dev. Sci.* 21, e12585
22. Tyas, S.L. and Pederson, L.L. (1998) Psychosocial factors related to adolescent smoking: a critical review of the literature. *Tob. Control.* 7, 409–420
23. Prentice, D.A. and Miller, D.T. (1993) Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *J. Pers. Soc. Psychol.* 64, 243
24. Yoeli, E. *et al.* (2013) Powering up with indirect reciprocity in a large-scale field experiment. *Proc. Natl. Acad. Sci. U. S. A.* 110, 10424–10429
25. Lacetera, N. and Macis, M. (2010) Social image concerns and prosocial behavior: field evidence from a nonlinear incentive scheme. *J. Econ. Behav. Organ.* 76, 225–237
26. Delmas, M.A. and Lessem, N. (2014) Saving power to conserve your reputation? The effectiveness of private versus public information. *J. Environ. Econ. Manag.* 67, 353–370
27. Yoeli, E. *et al.* (2019) Digital health support in treatment for tuberculosis. *N. Engl. J. Med.* 381, 986–987
28. Rogers, T. *et al.* (2016) Potential follow-up increases private contributions to public goods. *Proc. Natl. Acad. Sci.* 113, 5218–5220
29. Yoeli, E. and Rand, D. (2020) A checklist for prosocial messaging campaigns such as COVID-19 prevention appeals. *PsyArXiv* Published online April 17, 2020. https://psyarxiv.com/rg2x9
30. Kraft-Todd, G. *et al.* (2021) Public good messaging motivates the wealthy to reduce water consumption. *PsyArXiv* Published online January 2, 2021. https://psyarxiv.com/exz2a
31. Carattini, S. *et al.* (2023) Peer-to-peer solar and social rewards: evidence from a field experiment. *SSRN* Published online January 3, 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4316000
32. Gintis, H. *et al.* (2001) Costly signaling and cooperation. *J. Theor. Biol.* 213, 103–119
33. Alexander, R.D. (1987) *The Biology of Moral Systems*, Transaction Publishers
34. Boyd, R. and Richerson, P.J. (1989) The evolution of indirect reciprocity. *Soc. Networks* 11, 213–236
35. Roberts, G. (1998) Competitive altruism: from reciprocity to the handicap principle. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 265, 427–431
36. Zahavi, A. (1995) Altruism as a handicap: the limitations of kin selection and reciprocity. *J. Avian Biol.* 26, 1–3
37. Sylwester, K. and Roberts, G. (2013) Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evol. Hum. Behav.* 34, 201–206
38. Roberts, G. *et al.* (2021) The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *Philos. Trans. R. Soc. B* 376, 20200290
39. Sylwester, K. and Roberts, G. (2010) Cooperators benefit through reputation-based partner choice in economic games. *Biol. Lett.* 6, 659–662
40. Lyle, H.F. and Smith, E.A. (2014) The reputational and social network benefits of prosociality in an Andean community. *Proc. Natl. Acad. Sci.* 111, 4820–4825
41. Milinski, M. *et al.* (2002) Donors to charity gain in both indirect reciprocity and political reputation. *Proc. R. Soc. Lond. B Biol. Sci.* 269, 881–883
42. Barclay, P. (2006) Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* 27, 325–344
43. Nelissen, R. (2008) The price you pay: cost-dependent reputation effects of altruistic punishment. *Evol. Hum. Behav.* 29, 242–248
44. Raihani, N.J. and Bshary, R. (2015) Third-party punishers are rewarded – but third-party helpers even more so. *Evolution* 69, 993–1003
45. Horita, Y. (2010) Punishers may be chosen as providers but not as recipients. *Lett. Evol. Behav. Sci.* 1, 6–9
46. Kennedy, J.A. and Schweitzer, M.E. (2018) Building trust by tearing others down: when accusing others of unethical behavior engenders trust. *Organ. Behav. Hum. Decis. Process.* 149, 111–128
47. Hok, H. *et al.* (2019) When children treat condemnation as a signal: the costs and benefits of condemnation. *Child Dev.* 91, 1439–1455
48. Jordan, J.J. *et al.* (2017) Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychol. Sci.* 28, 356–368
49. Jordan, J.J. *et al.* (2016) Third-party punishment as a costly signal of trustworthiness. *Nature* 530, 473–476
50. Jordan, J.J. *et al.* (2016) Uncalculating cooperation is used to signal of trustworthiness. *Proc. Natl. Acad. Sci.* 113, 8658–8663
51. Barasch, A. *et al.* (2014) Selfish or selfless? On the signal value of emotion in altruistic behavior. *J. Pers. Soc. Psychol.* 107, 393–413
52. Levine, E.E. *et al.* (2017) Signaling emotion and reason in cooperation. *J. Exp. Psychol. Gen.* 147, 702–719
53. Critcher, C.R. *et al.* (2013) How quick decisions illuminate moral character. *Soc. Psychol. Personal. Sci.* 4, 308–315
54. Everett, J.A. *et al.* (2016) Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* 145, 772–787

55. Silver, I. and Shaw, A. (2022) When and why 'staying out of it' backfires in moral and political disagreements. *J. Exp. Psychol. Gen.* 151, 2542–2561
56. Zlatev, J.J. (2019) I may not agree with you, but I trust you: caring about social issues signals integrity. *Psychol. Sci.* 30, 880–892
57. Dorison, C.A. *et al.* (2022) Staying the course: decision makers who escalate commitment are trusted and trustworthy. *J. Exp. Psychol. Gen.* 151, 960
58. Dorison, C.A. (2022) A reputational perspective on rational framing effects. *Behav. Brain Sci.* 45, e226
59. Parsa, H.G. *et al.* (2015) Corporate social and environmental responsibility in services: will consumers pay for it? *J. Retail. Consum. Serv.* 22, 250–260
60. Burbano, V.C. (2016) Social responsibility messages and worker wage requirements: field experimental evidence from online labor marketplaces. *Organ. Sci.* 27, 1010–1028
61. Bhattacharya, C.B. *et al.* (2008) Using corporate social responsibility to win the war for talent. *MIT Sloan Manag. Rev.* 49, 49215
62. Turban, D.B. and Greening, D.W. (1997) Corporate social performance and organizational attractiveness to prospective employees. *Acad. Manag. J.* 40, 658–672
63. Bhattacharya, C.B. and Sen, S. (2004) Doing better at doing good: when, why, and how consumers respond to corporate social initiatives. *Calif. Manag. Rev.* 47, 9–24
64. Sen, S. *et al.* (2016) Corporate social responsibility: a consumer psychology perspective. *Curr. Opin. Psychol.* 10, 70–75
65. Portocarrero, F. and Burbano, V. (2022) Doing well by requiring employees to do good: field experimental evidence of the effects of a one-time, mandatory corporate social intervention on employees. *SSRN* Published online May 25, 2022. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4119943
66. Abraham, M. and Burbano, V. (2022) Congruence between leadership gender and organizational claims affects the gender composition of the applicant pool: field experimental evidence. *Organ. Sci.* 33, 393–413
67. Hilbe, C. *et al.* (2018) Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci.* 115, 12241–12246
68. De Freitas, J. *et al.* (2019) Common knowledge, coordination, and strategic mentalizing in human social life. *Proc. Natl. Acad. Sci.* 116, 13751–13758
69. Pinker, S. *et al.* (2008) The logic of indirect speech. *Proc. Natl. Acad. Sci.* 105, 833–838
70. Chakroff, A. *et al.* (2015) An indecent proposal: the dual functions of indirect speech. *Cogn. Sci.* 39, 199–211
71. Bursztyn, L. *et al.* (2020) Misperceived social norms: women working outside the home in Saudi Arabia. *Am. Econ. Rev.* 110, 2997–3029
72. Willer, R. *et al.* (2009) The false enforcement of unpopular norms. *Am. J. Sociol.* 115, 451–490
73. Boyd, R. and Richerson, P.J. (1992) Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethol. Sociobiol.* 13, 171–195
74. Jordan, J.J. and Kouchaki, M. (2021) Virtuous victims. *Sci. Adv.* 7, eabg5902
75. Chudek, M. and Henrich, J. (2011) Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends Cogn. Sci.* 15, 218–226
76. Handley, C. and Mathew, S. (2020) Human large-scale cooperation as a product of competition between cultural groups. *Nat. Commun.* 11, 702
77. Silver, I. *et al.* (2021) Selfless first movers and self-interested followers: order of entry signals purity of motive in pursuit of the greater good. *J. Consum. Psychol.* 31, 501–517
78. Carlson, R.W. and Zaki, J. (2018) Good deeds gone bad: lay theories of altruism and selfishness. *J. Exp. Soc. Psychol.* 75, 36–40
79. Gambetta, D. and Przepiorka, W. (2014) Natural and strategic generosity as signals of trustworthiness. *PLoS One* 9, e97533
80. Berman, J.Z. *et al.* (2014) The braggart's dilemma: on the social rewards and penalties of advertising prosocial behavior. *J. Int. Mark.* 52, 90–104
81. Heyman, G. *et al.* (2014) Children's sensitivity to ulterior motives when evaluating prosocial behavior. *Cogn. Sci.* 38, 683–700
82. De Freitas, J. *et al.* (2019) Maimonides' ladder: states of mutual knowledge and the perception of charitability. *J. Exp. Psychol. Gen.* 148, 158
83. Newman, G.E. and Cain, D.M. (2014) Tainted altruism: when doing some good is evaluated as worse than doing no good at all. *Psychol. Sci.* 25, 648–655
84. Kraft-Todd, G.T. *et al.* (2023) Virtue discounting: observability reduces moral actors' perceived virtue. *Open Mind* 1–22
85. Tosi, J. and Warmke, B. (2016) Moral grandstanding. *Philos Public Aff* 44, 197–217
86. Silver, I. *et al.* (2021) Inauthenticity aversion: moral reactance toward tainted actors, actions, and objects. *Consum. Psychol. Rev.* 4, 70–82
87. Hoffman, M. *et al.* (2015) Cooperate without looking: why we care what people think and not just what they do. *Proc. Natl. Acad. Sci.* 112, 1727–1732
88. Jordan, J. and Sommers, R. (2022) When does moral engagement risk triggering a hypocrisy penalty? *Curr. Opin. Psychol.* 47, 101404
89. Cha, S.E. and Edmondson, A.C. (2006) When values backfire: leadership, attribution, and disenchantment in a values-driven organization. *Leadersh. Q.* 17, 57–78
90. Effron, D.A. *et al.* (2015) Hypocrisy by association: when organizational membership increases condemnation for wrongdoing. *Organ. Behav. Hum. Decis. Process.* 130, 147–159
91. Wagner, T. *et al.* (2009) Corporate hypocrisy: overcoming the threat of inconsistent corporate social responsibility perceptions. *J. Mark.* 73, 77–91
92. Effron, D.A. *et al.* (2018) Hypocrisy and culture: failing to practice what you preach receives harsher interpersonal reactions in independent (vs. interdependent) cultures. *J. Exp. Soc. Psychol.* 76, 371–384
93. O'Connor, K. *et al.* (2020) Moral cleansing as hypocrisy: when private acts of charity make you feel better than you deserve. *J. Pers. Soc. Psychol.* 119, 540
94. Kreps, T.A. *et al.* (2017) Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *J. Pers. Soc. Psychol.* 113, 730
95. Hafenbrädl, S. and Waeger, D. (2021) The business case for CSR: a trump card against hypocrisy? *J. Bus. Res.* 129, 838–848
96. Graham, J. *et al.* (2015) When values and behavior conflict: moral pluralism and intrapersonal moral hypocrisy. *Soc. Personal. Psychol. Compass* 9, 158–170
97. Bolton, G.E. *et al.* (2005) Cooperation among strangers with limited information about reputation. *J. Public Econ.* 89, 1457–1468
98. Mathew, S. (2017) How the second-order free rider problem is solved in a small-scale society. *Am. Econ. Rev.* 107, 578–581
99. Martin, J.W. *et al.* (2019) When do we punish people who don't? *Cognition* 193, 104040
100. Miller, D.T. and Prentice, D.A. (2016) Changing norms to change behavior. *Annu. Rev. Psychol.* 67, 339–361
101. Jordan, J.J. and Rand, D. (2017) Third-party punishment as a costly signal of high continuation probabilities in repeated games. *J. Theor. Biol.* 421, 189–202
102. Michael, S. (1973) Job market signaling. *Q. J. Econ.* 87, 355–374
103. Zahavi, A. (1975) Mate selection – a selection for a handicap. *J. Theor. Biol.* 53, 205–214
104. Gintis, H. *et al.* (2003) Explaining altruistic behavior in humans. *Evol. Hum. Behav.* 24, 153–172
105. Barclay, P. (2013) Strategies for cooperation in biological markets, especially for humans. *Evol. Hum. Behav.* 34, 164–175
106. Nowak, M.A. and Sigmund, K. (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577
107. Panchanathan, K. and Boyd, R. (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* 224, 115–126
108. Ohtsuki, H. and Iwasa, Y. (2004) How should we define goodness? – reputation dynamics in indirect reciprocity. *J. Theor. Biol.* 231, 107–120

109. Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*, Palgrave Macmillan
110. Barclay, P. *et al.* (2021) Cooperating to show that you care: costly helping as an honest signal of fitness interdependence. *Philos. Trans. R. Soc. B* 376, 20200292
111. André, J.-B. (2010) The evolution of reciprocity: social types or social incentives? *Am. Nat.* 175, 197–210
112. Smith, E.A. and Bird, R.L.B. (2000) Turtle hunting and tombstone opening: public generosity as costly signaling. *Evol. Hum. Behav.* 21, 245–261
113. Harbaugh, W.T. (1998) What do donations buy?: a model of philanthropy based on prestige and warm glow. *J. Public Econ.* 67, 269–284
114. Bliege Bird, R. *et al.* (2018) The social significance of subtle signals. *Nat. Hum. Behav.* 2, 452–457
115. Dhaliwal, N.A. *et al.* (2022) Signaling benefits of partner choice decisions. *J. Exp. Psychol. Gen.* 151, 1446
116. Ohtsubo, Y. and Watanabe, E. (2009) Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evol. Hum. Behav.* 30, 114–123
117. Bolle, F. (2001) Why to buy your darling flowers: on cooperation and exploitation. *Theor. Decis.* 50, 1–28
118. Sozou, P.D. and Seymour, R.M. (2005) Costly but worthless gifts facilitate courtship. *Proc. R. Soc. B Biol. Sci.* 272, 1877–1884
119. Dhaliwal, N.A. *et al.* (2021) Reputational and cooperative benefits of third-party compensation. *Organ. Behav. Hum. Decis. Process.* 164, 27–51
120. Roberts, G. (2020) Honest signaling of cooperative intentions. *Behav. Ecol.* 31, 922–932
121. Frank, R.H. (1988) *Passions Within Reason: The Strategic Role of the Emotions*, WW Norton & Co
122. Manapat, M.L. *et al.* (2013) Information, irrationality, and the evolution of trust. *J. Econ. Behav. Organ.* 90, S57–S75
123. Levine, E.E. *et al.* (2018) Who is trustworthy? Predicting trustworthy intentions and behavior. *J. Pers. Soc. Psychol.* 115, 468
124. Hilbe, C. *et al.* (2015) Cooperate without looking in a non-repeated game. *Games* 6, 458–472
125. Roberts, G. (2015) Partner choice drives the evolution of cooperation via indirect reciprocity. *PLoS One* 10, e0129442
126. Leimar, O. and Hammerstein, P. (2001) Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B Biol. Sci.* 268, 745–753
127. Roberts, G. (2008) Evolution of direct and indirect reciprocity. *Proc. R. Soc. Lond. B Biol. Sci.* 275, 173–179
128. Fujimoto, Y. and Ohtsuki, H. (2023) Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment. *Proc. Natl. Acad. Sci.* 120, e2300544120
129. Murase, Y. *et al.* (2022) Social norms in indirect reciprocity with ternary reputations. *Sci. Rep.* 12, 455
130. Schmid, L. *et al.* (2023) Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nat. Commun.* 14, 2086
131. Panchanathan, K. and Boyd, R. (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432, 499–502
132. Ohtsuki, H. *et al.* (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* 457, 79–82
133. Henrich, J. and Boyd, R. (2001) Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* 208, 79–89
134. Henrich, J. (2004) Cultural group selection, coevolutionary processes and large-scale cooperation. *J. Econ. Behav. Organ.* 53, 3–35
135. Boyd, R. and Richerson, P.J. (2002) Group beneficial norms can spread rapidly in a structured population. *J. Theor. Biol.* 215, 287–296